DOCUMENT RESUME

ED 389 714 TM 024 194

AUTHOR Wainer, Howard; And Others
TITLE DIFferential Testlet Functioning Definitions and Detection. Program Statistics Research Technical Report No. 91-9.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-21
PUB DATE 91
NOTE 43p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Definitions; *Identification; *Item Bias; Item Response Theory; Scores; *Sex Differences; Statistical Analysis; Test Construction; Test Items; *Test Reliability
IDENTIFIERS *Likelihood Ratio Criterion; Scholastic Aptitude Test; *Testlets

ABSTRACT

It is sometimes sensible to think of the fundamental unit of test construction as being larger than an individual item. This unit, dubbed the testlet, must pass muster in the same way that items do. One criterion of a good item is the absence of differential item functioning (DIF). The item must function in the same way as all important subpopulations of examinees. In this paper, "testlet DIF" is defined, and a statistical methodology is provided to detect it. This methodology parallels the item response theory-based likelihood ratio procedures explored previously by D. Thissen, L. Steinberg, and H. Wainer (1988). The methodology is illustrated with analyses of data from a testlet-based experimental version of the Scholastic Aptitude Test. The illustration demonstrated that, in this instance, not modeling the testlet structure overestimated the reliability of the test by an amount equivalent to a test of doubled length. In addition, testlets made up of what appeared to be exemplary items exhibited significant sex DIF when testlets were considered "in toto." It was also found that testlets with modest DIF in both directions can still be fair at score and proficiency levels. (Contains 4 tables, 10 figures, a technical appendix, and 27 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

RR-91-21

ED 389 714

# DIFferential Testlet Functioning Definitions and Detection

Howard Wainer
Educational Testing Service

Stephen G. Sireci
Fordham University

David Thissen
University of North Carolina

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. COLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# PROGRAM STATISTICS RESEARCH

**TECHNICAL REPORT NO. 91-9**

Educational Testing Service
Princeton, New Jersey 08541



BEST COPY AVAILABLE

TM024194

Copyright © 1991. Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Reports requires the written consent of the author(s).

*Abstract*

It is sometimes sensible to think of the fundamental unit of test construction as being larger than an individual item. This unit, dubbed the testlet, must pass muster in the same way that items do. One criterion of a good item is the absence of DIF—the item must function in the same way in all important subpopulations of examinees. In this paper we define what we mean by *testlet DIF* and provide a statistical methodology to detect it. This methodology parallels the IRT-based likelihood ratio procedures explored previously by Thissen, Steinberg & Wainer (1988, in press). We illustrate this methodology with analyses of data from a testlet-based experimental version of the Scholastic Aptitude Test.

BEST COPY AVAILABLE

## DIFferential Testlet Functioning
## Definitions and Detection

It is often natural to think of the fungible unit of a test as a testlet: an interrelated and integrated group of items, always presented as a single unit (Wainer & Kiely, 1987). Historically tests of skills such as reading comprehension have been constructed of testlets: a passage followed by a number of interrelated questions (Thissen, Steinberg & Mooney, 1989). Recent trends in test construction (Resnick, 1987; National Council of Teachers of Mathematics, 1989) emphasize a global view in the assessment of proficiency. This trend toward focusing tests on a less microscopic level than the item indicates a rich future for the use of testlets.

In parallel with this call for tests with greater construct validity has been a renewed emphasis on issues of test fairness. One aspect of fairness is the insistence that test items not function differentially for individuals of the same proficiency, regardless of their group membership. "No DIFferential item functioning" is now a general desideratum; the area of study surrounding this has been defined formally and dubbed DIF. A set of statistically rigorous and efficient procedures have been developed to detect and measure DIF. These generally fall into one of two classes; they are either based on latent variables (Thissen, Steinberg & Wainer, 1988, in press) or observed score (Holland & Thayer, 1988; Dorans & Holland, in press).

Procedures for DIF studies have traditionally focused on the item; indeed "item" is sometimes thought of as DIF's middle name! Yet, if future tests will be based on testlets ought we not generalize DIF procedures to suit this broader construct? The point of this paper is to argue for precisely such a generalization and to provide two allied methods for accomplishing it. Because this article is statistical we will not address the issues surrounding what one does with a testlet that is found to contain DIF. These issues are typically nonstatistical in nature involving decisions made on the basis of content and practicality. We leave such a discussion to other accounts.

## Testlet DIF—An Inevitable Concept

The determination of DIF at the testlet level has three advantages over confining the investigation to the item. It allows:

1. the analysis model to match the test construction,
2. DIF cancellation through balancing,
3. the uncovering of DIF that, because of its size, evades detection at the item level but can become visible with some aggregation.

### *Matching the Model to the Test*

If a set of items were built to be administered as a unit, it is important that they be analyzed that way. There are a variety of reasons for analyzing them as a unit, but underlying them all is that if one does not, one is likely to get the wrong answer. In the example described in a subsequent section, a four testlet test consisting of 45 separate items yields a reliability of .87 if calculated using traditional methods assuming 45 independent items. If one calculates reliability taking the within-testlet dependencies into account, the test's reliability is shown to be .76. These are quite different—note that Spearman-Brown (Gulliksen, 1950, p. 78) indicates that we would need to double the test length to yield such a gain in reliability [see Sireci, Thissen & Wainer (1990) for more details on this aspect]. Other calculations (i.e., validity and information) are affected as well.

### *DIF Cancellation*

Roznowski (1988), among others, has pointed out that because decisions are made at the scale or test level, DIF at the item level may have only limited importance. Therefore it it sensible to consider an aggregate measure of DIF. Small amounts of item DIF that cancel within the testlet would seem, under this argument, to yield a perfectly acceptable test construction unit. This is of critical importance in adaptive testing, less so with fixed format tests.

Humphreys (1962, 1970, 1981, 1986) has long argued that it is both inadvisable and difficult—very likely impossible—to try to construct a test of strictly unidimensional items. He suggests that to do so would be to construct a test that is sterile and too far abstracted from what

would be commonly encountered to be worthwhile. He recommends the use of content rich (i.e., possibly multidimensional) items, and since multidimensionality is what causes DIF, we ought to control it by balancing across items. We agree with this. But balancing is not a trivial task. Surely such balancing needs to be done within content area and across the entire test. For example, it would be unfortunate if the items that favored one group were all at the end of the test. The concept of a testlet suggests itself naturally. Build the test out of testlets and ensure that there is no DIF at the testlet level. Lewis & Sheehan (1990) have shown that building a mastery test of parallel-form testlets provides a graceful solution to a set of thorny problems.

Cancellation of DIF could be accomplished in an adaptive testing situation without using testlets. However, it would involve accumulating DIF statistics of the items as they are to be administered and ensuring that the accumulation was zero when the test halted. This is almost surely possible without testlets, but it would certainly add a further burden to the item selection algorithm and item pool. Providing DIF-balanced testlets as the unit of test construction seems a much simpler strategy.

A final argument in support of examining DIF at the testlet level derives from the consideration of testlets that cannot easily be decomposed into items. For example, consider a multistep mathematics problem in which students get credit for each part successfully completed. Does it make sense to say that parts of such a testlet contain "positive subtraction DIF" and then "negative multiplication DIF"? Of course not. Instead we must concentrate on the DIF of the problem as a whole. In some sense we do this now when we test an item's DIF. We do not record intermediate results and so do not know to what extent there is DIF on the component tasks required to complete the item. All we concern ourselves with is the final result.

It should be emphasized that by cancel out we mean something quite specific. We mean that there will be no DIF at every score level within the testlet. Exactly how we operationalize this goal and what it means will be explicated and illustrated in the next sections.

*Increased Sensitivity of Detection*

It is possible (and, as we will demonstrate, even likely) to construct a testlet of items with no detectable item DIF, yet the testlet in the aggregate does have DIF. The increased statistical power of dealing with DIF at the testlet level provides us with another tool to insure fairness. This will be especially useful for those focal groups that are relatively rare in the examinee population and so are not likely to provide large samples during item pretesting.

**Testlet DIF detection—One Model, Two Methods**

The polytomous IRT model we used was developed by Bock (1972). The basic notion is to fit the model to the data assuming that all testlets have the same parameters (no DIF) in the two populations of interest (Reference and Focal). We then fit the same model to the data allowing one testlet to have different parameters in each population (DIF) and compare the likelihood under each of the two situations. If the more general model does not yield a significant increase in the quality of the fit we conclude that the extra generality was not needed and that the testlet in question has no DIF. This procedure was applied in the study of DIF by Thissen et al. (1988) using a more traditional dichotomous IRT model. Thissen et al. (1989) used Bock's polytomous model to fit testlets. Our testlet approach to DIF is almost exactly the one reported by Thissen, et al., (in press) when we used the multiple choice model (Thissen & Steinberg, 1984) to examine differential alternative functioning (DAF). The step from DAF to testlet DIF is a small one.

*Bock's 1972 Model*

Suppose we have $J$ testlets, indexed by $j$, where $j = 1, 2, ..., J$. On each testlet there are $m_j$ questions, so that for the $j$th testlet there is the possibility for the polytomous response $x_j = 0, 1, 2, ..., m_j$. The statistical testlet scoring model posits a single underlying (and unobserved) dimension that we call latent proficiency, and denote $\theta$. The model then represents the probability of obtaining any particular score as a function of proficiency. For each testlet there is a set of functions, one for each response category. These functions are sometimes called item

characteristic curves (Lord & Novick, 1968), item operating curves (Samejima, 1969) or trace lines (Thissen et al., 1989). We shall follow Thissen et al.'s (1989) notation and nomenclature.

The trace line for score $x = 0, 1, ..., m_j$, for testlet $j$ is

$$T_{jx}(\theta) = \frac{\exp[a_{jx}\theta + c_{jx}]}{\sum_{k=0}^{m} \exp[a_{jk}\theta + c_{jk}]} \quad ; \tag{1}$$

where the $\{a_k, c_k\}j$, $k = 0, 1, ..., m_j$ are the item category parameters that characterize the shape of the individual response trace lines. The $a_k$s are analogous to discriminations; the $c_k$s analogous to intercepts. The model is not fully identified, and so we need to impose some additional constraints. It is convenient to insist that the sum of each of the sets of parameters equal zero, i.e.

$$\sum_{k=0}^{m_j} a_{jk} = \sum_{k=0}^{m_j} c_{jk} = 0 \quad .$$

In this context, we reparameterize the model using centered polynomials of the associated scores to represent the category-to-category change in the $a_k$s and the $c_k$s:

$$a_{jk} = \sum_{p=1}^{P} \alpha_{jp} \left(k - \frac{m_j}{2}\right)^p \tag{2}$$

and

$$c_{jk} = \sum_{p=1}^{P} \gamma_{jp} \left(k - \frac{m_j}{2}\right)^p \tag{3}$$

where the parameters $\{\alpha_p, \gamma_p\}j$, $p = 1, 2, ...P$, for $P \leq m_j$ are the free parameters to be estimated from the data. The polynomial representation has, in the past, saved degrees of freedom with no significant loss of accuracy. It also provides a check on the fit of the model when the categories are ordered. Although this model was developed for the nominal case it can be used for ordered categories. If the categories are ordered the $a$'s must be monotonically ordered (see the appendix

for proof). As we show in the next section the polynomial representation in this application saves degrees of freedom and indicates that the model provides a good representation of the data.

This version of Bock's model uses raw score within testlet as the carrier of information. While it is possible that more information would be obtained by taking into account the pattern of responses within each testlet we felt that this simplification is appropriate for an initial foray into testlet DIF. Moreover, basing a test scoring algorithm on number right seems amply supported by general practice, especially as a first step.

In previous work, this model was fitted to a 4-passage, 22 item test of reading comprehension by Thissen et al. (1989), with $m_j$=(7, 4, 3, 8). The analysis followed an item factor analysis (Bock, Gibbons & Muraki, 1988) that showed that a multifactor structure existed. The (at least) 4-factor structure found among these 22 items made the unidimensional assumption (conditional independence) of traditional IRT models untenable. After considering the test as four testlets and fitting Bock's nominal response model to the data generated by the almost 4,000 examinees, they compared the results obtained with what would have been the case if they had ignored the lack of conditional independence and merely fit a standard IRT model. They found two things: First that there seemed to be a slightly greater validity of the testlet derived scores when correlated with an external criterion. Second, the test information function yielded by the traditional analysis was much too high. This was caused by this model's not being able to deal with the excess intra-passage correlations among the items (excess after conditioning on $\theta$). The testlet approach thus provided a more accurate estimate of the accuracy of the assessment. Through an obvious generalization, this same approach can be used to study Testlet DIF.

*Method 1: Internal Criterion*

The basic data matrix of score patterns is shown in Table 1. In this example, there are four testlets with 10 possible scores levels each [$m_j$=(10, 10, 10, 10)]; there are a maximum of $10^4$ rows. In practice there will be far fewer rows since many possible response patterns will not appear. The analysis follows what is done in item DIF situations: fitting one model allowing

different values for the parameters of the studied testlet for the two groups and then comparing the -2loglikelihoods of that model with others that restrict the two groups' estimates in a variety of ways. Stratification/conditioning is done on $\theta$ estimated for both groups simultaneously.

---

Insert Table 1 About Here

---

This method uses the test itself, including the studied testlet, to calculate the matching criterion. The question about whether or not to include the studied item has been carefully explored (Holland & Thayer, 1988) who showed for the Rasch model (the binary analog of this model) that not including the studied item in the criterion yields statistical bias under the null hypothesis. This was explored further by Zwick (1990) who confirmed this result for the Rasch model, but not generally for other IRT models.

Using this method requires first fitting a completely unrestricted model—estimating all of the $a_k$ s and $c_k$s separately for both the reference and the focal groups. Next restricted versions of this model are estimated by approximating the values of the parameters as polynomial functions of score category (equations 2 and 3). When an acceptably fitting parsimonious model is derived we note the value of -2loglikelihood (asymptotically $\chi^2$) for that model and then sequentially restrict the parameters for one testlet at a time to be equal across the two groups. We subtract the -2loglikelihood from the restricted model from the unrestricted and, remembering that the difference between two $\chi^2$ statistics is also $\chi^2$, we test that difference for significance; the number of degrees of freedom of the statistical test is equal to the number of parameters restricted. If it is not significant we conclude that the extra flexibility gained by allowing different parameters for the focal and reference groups is not required—there is no DIF. If it is significant we can further isolate where the DIF is located.

Eventually one arrives at a determination of the most parsimonious representation. Interpreting the character of this representation allows us to detect testlet DIF. This is computationally expensive, with the cost of each run essentially linear in the number of response

patterns observed. Of course this cost is small relative to the cost of not detecting testlet DIF when it is there. The cost can be controlled substantially by reducing the number of possible response patterns. One way to do this is explored in the next section.

*Method 2: External Criterion*

The basic data matrix of score patterns is shown in Table 2. There is a matrix like this for both the Reference group ($G=R$) and the Focal group ($G=F$). For convenience this example uses a six item anchor yielding $2^6$ or $N = 64$ possible matching levels.

---

Insert Table 2 About Here

---

This method uses an external criterion as the matching variable. This has been recommended as the practice of choice when a suitable external measure is available (Angoff, 1982, pp. 112-113; Thissen et al., in press). It cleanly avoids the issues surrounding what to do with the studied item when the matching criterion is internal, as well as arguments of circularity. Of practical importance, the analysis focuses on a matrix $640 \times 2$: only 1,280 cells. This allows many items to be examined at only a modest cost in computer time. Contrast this with the parallel task utilizing an internal anchor that has 2-by-$10^4$ or 20,000 cells. The former analyses can be easily accommodated on a microcomputer; the latter is more comfortable on something larger, faster and more expensive.

The strategy for accomplishing this analysis is quite similar to that described in the previous section. But there is one important extra step—the choice of the criterion items. We will not deal with the substantive aspects of that choice in this section; instead we focus on the psychometric characteristics used in the choice. The criterion items should: *a.* be strongly related to the same underlying characteristic that is being measured by the testlets, *b.* have steep slopes, *c.* have their difficulties span the range of proficiency of the individuals taking the test, and *d.* have no DIF. How many items are required? We have been successful with as few as three, but a

more conservative stance (yielding protection against one of these items behaving poorly) would use five or six or seven. We chose six in the example reported here—it worked very well indeed.

After choosing these special items [Thissen et al. (in press) called these the designated anchor], each testlet takes its turn as the "studied testlet." A saturated model is fitted, followed by suitably restricted ones. When the likelihood ratio indicates that restricting the testlet's parameters to be equal across the two groups does not worsen the fit we conclude no DIF and move on to the next testlet. If it is significant we continue our explorations to try to isolate the specific parameters that characterize the DIF.

Our experience with this methodology indicates that we obtain essentially the same results as with the more costly internal method. In the example described here the computing time was about one third that used with the internal method. This is as expected since the size of the matrices used with the external criterion are about one-third that using the internal criterion. With a larger sample of individuals, and hence more different response patterns, the difference would be more dramatic still.

**Testlet DIF Applied—The NPP-V**

The data analyzed here were part of the Spring, 1989 field testing of the New Possibilities Prototype test (NPP), an experimental version of the Scholastic Aptitude Test (SAT). This field testing represents an ongoing collaborative effort of the College Board and ETS designed to investigate possible enhancements to the current SAT. The verbal section of the NPP, the NPP-V, includes longer reading passages than the SAT, and has more items associated with each passage. The form of the NPP-V analyzed here consists of 75 multiple-choice items, 45 of which correspond to four long reading passages. These reading passages have 12, 13, 10 and 10 corresponding items respectively. We shall henceforth refer to these four passages as Testlets I, II, III, and IV. A more complete description of the NPP is not currently available, but will be within the year. The analyses were based 4,028 high school students: 2,216 females and 1,812 males.

*Analysis Preliminaries*

All analyses were done using MULTILOG Version 6.0 (Thissen, 1990); it allows the mixing of item types within the same analysis; that is crucial for the use of an external anchor of dichotomous items. It also allows the imposition of equality constraints; that is necessary to obtain the likelihood of restricted (no DIF) models.

The maximum number of categories that the current version of MULTILOG allows for any polytomous model is ten. This limit required that we collapse some of the response categories in Testlets I and II. Since categories with very few entries provides poor parameter estimates we found that little power was lost and, indeed some stability was gained,[2] by combining some extreme score categories. Testlet I's 12 categories were reduced to 10 by combining score groups 0 and 1 into a new group labeled '0' and categories 11 and 12 into a single category labeled '9.' Testlet II's 13 score categories were similarly reduced by combining the three lowest (0, 1, 2) into category '0' and two highest into category '9.'

Previous experience (Thissen et al., 1989) has shown that trace lines for essentially chance scores are sufficiently similar to one another so that they can be combined with no loss of information. These are all five choice items and so we would expect chance performance on such testlets to yield scores around 2. The number of individuals in the highest categories were sufficiently small so that the judicious melding of those score categories would yield nothing but statistical stability. Thus we felt that this accommodation to the limits of the current version of MULTILOG would not influence our results.

*Results of Method 1: Internal Anchor*

The analysis began by fitting a completely unconstrained model to the data. This allowed each testlet to be fitted separately by sex and the polynomials described in equations (2) and (3) were of ninth degree. We subsequently found that for the four testlets fitted we never needed greater than third degree polynomials, and that often linear or quadratic functions gave wonderful fits.

Shown in Figure 1 are the the fitted (line) and actual (points) values for the $a_k$ values for Testlet I for males. We reproduce these here to show the closeness that (in this case) a quadratic approximation has to the actual data (this is the worst fitting set of parameters in this study). In Figure 2 are shown the fitted and actual values for the $c_k$s for Testlet I obtained from the male examinees. Once again this is the worst fitting polynomial that we found. Once again the constrained values are depicted with the fitted line, the unconstrained by the plotted points.

---

Insert Figures 1 & 2 About Here

---

After determining the proper level of generality for the polynomial representation for the parameters of each of the testlets (examined separately for men and women) we began fitting a sequence of hierarchically nested models. We started with a completely unconstrained model (commonly termed fully saturated) in which each testlet had different parameters for males and females. Next we fit a completely constrained model in which the parameters for each testlet administered to males were constrained to be equal to the corresponding parameters of that testlet administered to females. We then moved from the constrained to the unconstrained model in directed steps. The results are shown in Table 3 and summarized graphically in Figure 3.

---

Insert Figure 3 & Table 3 About Here

---

We can quickly see that the No DIF model can be rejected out of hand. The next sequence of four models tests whether the DIF can be isolated within a single testlet. The answer is no, but we get some useful information about what is going on. There are major decreases in misfit when Testlets I and II are allowed to show DIF, but the allowing Testlets III and IV to have separately estimated parameters by sex yields no increase in the quality of the fit. It appears that it is likely that the DIF is located in Testlets I and II. The next row of Table 3 shows that when we fit a model that restricts Testlets III and IV to be equal in both groups but allows

separate estimation in I and II the fit is not significantly different than the unconstrained model. Plotting the parameters separately estimated for Testlets I and II suggested that both the discriminations and thresholds for Testlet II were quite different for the two sexes, but for Testlet I only the $c_k$s seemed to be different. Figures 4 and 5 show these plots for Testlet I.

---

Insert Figures 4 and 5 About Here

---

The information in Figure 4 led us to constrain the $a_k$s in Testlet I and thus we arrive at the final model that indicates DIF in Testlet II in both discrimination and location parameters, in Testlet I in only location, and no DIF in either Testlet III or IV. Now that we have located the DIF, it remains that we try to understand it. Plots of parameters of a polytomous IRT model are not always easy to figure out. The next step is to examine the trace lines associated with these parameters.

In the interests of parsimonious presentation we will not reproduce here the trace lines for all of the testlets; instead we will focus on Testlet I. We do this to illustrate a variety of points; key among them is the size of DIF detectable with this methodology and this sample size. In the upper and lower panels of Figure 6 are the trace lines for Testlet I for males and females respectively. They look remarkably similar; however, the trace lines for the males are shifted to the left, relative to the female trace lines, for the higher testlet scores. The location of the shift in the trace lines shows where there is DIF and the amount of shift indicates the amount of DIF. Evaluating the size of the DIF requires weighting the differences by the proficiency distribution of the focal group (Wainer, in press). The DIF is difficult to see in Figure 6; fortunately there is another way to examine the result.

---

Insert Figure 6 About Here

---

Each of the trace lines in Figure 6 indicates the conditional probability of an individual being in that score group [$P(x = k \mid \theta)$ for $k = 0, 1, ..., 9$]. Plotting the expected conditional score group [$E(x \mid \theta) = \Sigma [x P(x \mid \theta)]$] reduces the sheaf of ten curves for each sex to a single function. The expected score group is very close in both form and spirit to Lord's (1980) recommendation regarding the use of expected true score. Shown in Figure 7 is a plot of the expected score groups for males and females. The direction of the advantage is clear. If we subtract the females' curve from the males' we obtain a clear depiction of the size of the DIF (see Figure 8). From this plot we see that the maximum advantage is about a half point on a ten point scale (about 5%).

---

Insert Figures 7 and 8 About Here

---

Before concluding this section let us examine the size and direction of the DIF found in Testlet II using plots of expected score category. The difference plot for Testlet II is shown in Figure 9. We see immediately that this time the DIF is in favor of females with the advantage disappearing at higher proficiency levels. The effect of unequal discriminations ($a_k$s) is apparent. We also note (in Figure 9) that the maximal advantage to females is about one point (out of ten), and it is centered at about the center of the proficiency distribution. This is roughly twice the DIF seen in Testlet I.

---

Insert Figure 9 About Here

---

The size and direction of the testlet DIF detected provides a sense of the statistical power of this methodology. Testlets III and IV had no detectable DIF. By that we mean any DIF that might exist within those two testlets was smaller than that shown here. It should be emphasized that there were only a few items in any of the testlets that showed significant DIF when screened individually [3]. Thus, examining entire testlets for DIF as a whole has provided us with a tool with increased sensitivity. However the cost of this increased statistical power has been a

substantial increase in the conceptual and computational complexity. In the next section we show how a much simpler methodology gets us essentially identical results.

*Results of Method 2: External Anchor*

When we use an internal anchor the basic data matrix is potentially very large indeed; $10^4 \times 2$ for 4 testlets with ten score categories. The size of the analysis problem goes up exponentially with the number of testlets—$10^5$ for five testlets, $10^6$ for six, etc. This can be controlled and sharply reduced through the use of a fixed external anchor. In this case we chose six dichotomous items from among the 30 multiple choice items that were also on the NPP-V. We chose these items very carefully indeed. They were the items with the lowest DIF (measured using the Mantel-Haenszel statistic) that spanned the range of the proficiency distribution. We also tried to choose items that had good discrimination.

These six anchor items were fitted with the three parameter logistic IRT model (3-PL) in the course of the DIF analyses. In Figure 10 are the estimated parameters for these items and plots of their trace lines.

---

Insert Figure 10 About Here

---

Once the anchor items were chosen we followed the procedure described earlier. This required appending each testlet in turn as the "studied testlet" to the six-item anchor. Then we fitted an unconstrained model allowing the testlet to have different parameters for the focal and reference group, and a model in which the testlet's parameters were constrained to be equal in the two groups. Once again we looked at the likelihood ratio and if there was no significant increase in fit with the relaxation of the equality constraints we concluded that there was no DIF. The results of these analyses are summarized in Table 4.

---

Insert Table 4 About Here

---

It is clear that the conclusions we would draw from the results shown in Table 4 are the same as those drawn from the internal analyses; this adds empirical support to the practical reasons for using a short external anchor of multiple choice items to stratify the examinee population. Testlet I shows DIF only in the threshold parameters (the $c_k$s)—note that the likelihood ratio $\chi^2$ comparing a model restricting just $a_k$s with one with no restrictions is 2 on 2 degrees of freedom. Testlet II shows DIF; Testlets III and IV show no DIF. Plots of the trace lines for the testlets estimated within the context of an external anchor are virtually identical to those obtained with the far more computationally intensive internal anchoring procedures. Each estimation run here requires analysis of a $64 \times 10 \times 2$ table (1,280 cells). Two runs are required for each testlet. The internal anchor is far more complex. As the number of testlets increases computing time using an external anchor increases linearly, whereas using an internal anchor it increases exponentially. Our experience so far suggests that if a good anchor can be constructed one would be foolish not to use it.

Restrictions surrounding security of operational test forms precludes us from any extensive discussion of the content of the testlets analyzed in this paper, however the passage associated with testlet II (DIF favoring women) was an extended description of someone's visit to their grandmother. Testlet IV involved excerpts from speeches by Pericles and Abraham Lincoln.

**Testlet DIF whither—A discussion of what's next**

It is to the benefit of large testing organizations to look for DIF and not find any. In statistical terminology not finding DIF means being unable to reject the null hypothesis. That is we assume that there is no DIF, and after considering the evidence decide that we cannot reject that hypothesis. It is easy to accept the hypothesis that there is no DIF. To accomplish this one merely has to run poor studies with smallish sample sizes and use weak statistical models. Thus to be credible, a finding of no DIF must be accompanied by a careful study with as large a

sample size as could be found. It must also use the strongest statistical model available to analyze these data; where by strongest we mean most efficient.

The history of DIF procedures, described by Angoff (in press), illustrates how statistical methods were initially developed to match heuristic ideas about what ought to be measured. This was, properly, the most important initial concern, with niceties such as statistical power being left for later. In the past few years two classes of powerful models for detecting and measuring DIF became available. Dorans & Holland (in press) provide a thorough description of two quite similar procedures (one based on standardization and the other on the Mantel-Haenszel statistic). These methods are nonparametric in that they do not attempt to model response likelihoods. Both methods are statistically efficient and inexpensive to compute.

Thissen et al. (in press) describe methods that utilize a likelihood ratio of two models to detect DIF. Statistical theory predicts that these methods are asymptotically optimal when the IRT model that is assumed to underlie the individual item responses is appropriate. In this chapter the authors demonstrate how the methodology generalizes easily to study patterns of differential response among the item's distractors. This generalization is achieved through the use of a polytomous IRT model and results in what the authors call a methodology for studying DAF. This powerful new tool is shown to be helpful in diagnosing the misfunctioning of an item after DIF has been detected.

In the current presentation we have generalized DAF procedures to allow the detection of Testlet DIF. We have shown that this generalization accomplishes a variety of worthy goals. We showed that:

1. It characterizes the statistical character of the test more accurately than is the case with any model that does not acknowledge the clustered structure of the test's items. We illustrated this when we pointed out that by not modeling the testlet structure the reliability of the test was overestimated by an amount equivalent to a test of doubled length.

2. Testlets made up of what appeared to be exemplary items (both Testlets I and II), exhibited significant sex DIF when the testlets were considered in toto. This increased statistical

power is especially important when we study the suitability of newly developed items for subpopulations of examinees who show up only seldom in test samples.

3. Testlets constituted of items with modest DIF in both directions can still be fair at all score and proficiency levels (Testlet IV).

We believe that since the current weltanschauung points test development toward tests composed of larger tasks it is well that we have the statistical tools to properly deal with such tests. The concept of the testlet and the associated psychometrics is a big step in that direction.

We recognize that procedures based upon the fitting of hierarchical IRT models and the examination of likelihood ratios does not meld well with the economic stringencies of mass testing. Imagine the resources required for a detailed examination of the thousands of items required for an adaptive item pool! It would surely be better if something more computationally parsimonious could be found. Paul Holland often promoted the Mantel-Haenszel by exclaiming "10¢ an item!" Perhaps for a 13 item testlet he would be content with achieving the goal of "$1.30 a testlet." Using an internal anchor does not approach this goal, although it does allow a level of detail in the investigation that has not been approached yet with other methods. The external anchor methodology is much more practical while sacrificing little or none of the power of the internal method. It also illustrates the single greatest strength of IRT-based methods—it can stratify individuals on a short anchor (Bock, in press). Nonmodel-based methods like those utilizing the Mantel-Haenszel or standardization stratify examinees on their raw score. This works fine when a test is long enough to do this reliably. But on short tests reliable stratification requires utilizing information from response patterns. Some IRT models do this and so yield the accuracy of result we illustrate here.

We believe that testlet-based generalizations of Mantel-Haenszel can be usefully applied. For example one obvious generalization would stratify individuals by score and then within score stratum calculate the contingency table of testlet score-by-group membership. Current usage with dichotomous items yields a $2 \times 2$ matrix (correct-incorrect $\times$ Focal-Reference); this generalization would yield an $m_j \times 2$ matrix. The statistic would then (as now) be calculated

under the hypothesis of no interaction. Of course this sort of generalization could be used with dichotomous items but with several groups—test for DIF in all (say $n$) focal groups at once! Why hasn't it been used this way? The answer relates to the statistical fact that one degree of freedom tests are the most powerful. Thus we achieve a more sensitive detection instrument if we do a series of one degree of freedom tests rather than a single ($n$ - 1) degree of freedom test. To match this attitude in using the Mantel-Haenszel to detect testlet DIF one might want to collapse score categories to just two (perhaps above and below average). Then grind on with the usual Mantel-Haenszel procedure. Contrast this more extreme approach with our practice of collapsing to ten categories. We suspect that we could obtain somewhat more power but at a cost of understanding exactly where the problem lies.

We considered these arguments in our development of the methodology presented here. We believe that we have arrived at a sensible compromise between power and delicacy. Anyone doubting this should consider the size of the differences that were uncovered as being statistically significant (see Figure 7). The samples we have used are realistic for most practical situations to reliably detect rather small amounts of DIF using an anchor of only six items. We believe that this represents a method of sufficient power for most applications.

## References


Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Eds.) *Handbook of methods for detecting item bias* (pp. 96-116) Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H. (in press). Perspectives on the theory and application of differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning: Theory and Practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bock, R. D. (in press). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning: Theory and Practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Dorans, N. J., & Holland, P. W. (in press). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning: Theory and Practice* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (Pps. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475-483.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.) *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23-32). Seattle: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.) *Intelligence and Learning* (pp. 87-102). New York: Plenum Press.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327-333.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* xxx-xxx.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Resnick, L.B. (1987). *Education and learning to think.* Committee on Mathematics, Science, and Technology Education, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Roznowski, M. (1988). Review of *Test Validity. Journal of Educational Measurement, 25,* 357-361.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 4,* Part 2, No. 17.

Sireci, S. G., Thissen, D., & Wainer, H. (1990). On the reliability of testlet-based tests. Manuscript submitted for publication.

Thissen, D. (1990). *MULTILOG (version 6.0) user's guide.* Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items, *Psychometrika, 49,* 501-519.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (Pps. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (in press). Detection of Differential Item Functioning using the Parameters of Item Response Models. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning: Theory and Practice* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (in press). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning: Theory and Practice* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-198.

## Technical Appendix[4]

On an item that is scored in an ordered scale, 1, 2, ..., $m$, we would like the odds of being in a higher score category to be greater for an examinee of greater proficiency than for one with less.

Stated symbolically,

$$\frac{P(x = j \mid \theta = \theta_1)}{P(x = k \mid \theta = \theta_1)} > \frac{P(x = j \mid \theta = \theta_2)}{P(x = k \mid \theta = \theta_2)} \qquad (A1)$$

where $x$ is the observed score, $j > k$, $\theta$ is proficiency and $\theta_1 > \theta_2$.

Using shorthand notation $P(x_j \mid \theta_1)$ to mean $P(x = j \mid \theta = \theta_1)$ and rearranging allows us to rewrite inequality (A1) as

$$\frac{P(x_j \mid \theta_1)\, P(x_j \mid \theta_2)}{P(x_k \mid \theta_1)\, P(x_k \mid \theta_2)} > 1 . \qquad (A2)$$

Taking logs yields

$$\ln [P(x_j \mid \theta_1)] + \ln [P(x_k \mid \theta_2)] - \ln [P(x_j \mid \theta_2)] - \ln [P(x_k \mid \theta_1)] > 0. \qquad (A3)$$

If we model the probabilities with Bock's (1972) formulation for a categorical model (Equations 1 and 2) we find that

$$\ln [P(x_j \mid \theta_1)] = a_j \theta + c_j - \ln [denominator] .$$

After substituting this in inequality (3), we find that the denominators cancel out and we are left with

$$(a_j \theta_1 + c_j) + (a_k \theta_2 + c_k) - (a_j \theta_2 + c_j) - (a_k \theta_1 + c_k) > 0 \qquad (A4)$$

Rearranging and cancelling yields

$$a_j - a_k > 0$$

or

$$a_j > a_k \quad \text{for } j > k \qquad (A5)$$

This tells us that to accomplish our goal requires an ordering of the $a$ parameters.

Thus, the practice of fitting a monotone function to initially estimated $a$'s not only provides a more parsimonious model, but insures that Bock's nominal model yields satisfactory results for ordered categories of scoring.

## Footnotes

[1]This work was supported by the Educational Testing Service through funding from the Program Research Planning Council and the New Possibilities Project. We would like to express our gratitude for this help. The work of the second author was done while he was an ETS Summer Predoctoral Fellow. The data reported on were made available to us by Nancy Feryok, John Fremer and Ida Lawrence. We would like to express our gratitude to them as well as to: Bill Angoff, Neil Dorans, Paul Holland, Charles Lewis, Robert Mislevy, Alicia Schmitt, William Stout and Michael Zieky for their advice and comments on the work as it progressed. The responsibility for any errors that remain is ours.

[2]Coefficient $\alpha$ was higher for summed scores with the extreme score categories collapsed than it was with the original data.

[3] Actually there were a couple of items in Testlet IV that demonstrated a modest amount of DIF. However these items were counterbalanced by others that showed small DIF in the other direction. As we have shown, this counterbalancing, whether intentional or not, was effective in yielding a testlet that has no significant DIF in any score category at any value of proficiency.

[4]Our thanks to Paul Holland and Charles Lewis who, respectively, pointed out the mathematical relationship described here and its importance.

TABLE 1
Arrangement of the Data for the Internal Analyses

| Testlet Score Pattern | | | | Total | Frequencies | |
|---|---|---|---|---|---|---|
| I | II | III | IV | Score | Reference | Focal |
| 0 | 0 | 0 | 0 | 0 | $f_{R1}$ | $f_{F1}$ |
| 0 | 0 | 0 | 1 | 1 | $f_{R2}$ | $f_{F2}$ |
| 0 | 0 | 0 | 2 | 2 | $f_{R3}$ | $f_{F3}$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| $s_I$ | $s_{II}$ | $s_{III}$ | $s_{IV}$ | $\sum_i s_j$ | $f_{Ri}$ | $f_{Fi}$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 9 | 9 | 9 | 9 | 36 | $f_{RN}$ | $f_{FN}$ |

TABLE 2
Arrangement of the Data for the External Analyses

| Criterion items | | | | | | Testlet score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 0 | 0 | 0 | 0 | 0 | $f_{G10}$ | $f_{G11}$ | $f_{G12}$ | $f_{G13}$ | $f_{G14}$ | $f_{G15}$ | $f_{G16}$ | | | |
| 0 | 0 | 0 | 0 | 0 | 1 | $f_{G20}$ | $f_{G21}$ | $f_{G22}$ | $f_{G23}$ | $f_{G24}$ | $f_{G25}$ | $f_{G26}$ | | | |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $f_{G0}$ | $f_{G1}$ | $f_{G2}$ | $f_{G3}$ | $f_{G4}$ | $f_{G5}$ | $f_{G6}$ | | | |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 1 | 1 | 1 | 1 | 1 | 1 | $f_{GN0}$ | $f_{GN1}$ | $f_{GN2}$ | $f_{GN3}$ | $f_{GN4}$ | $f_{GN5}$ | $f_{GN6}$ | | | |

TABLE 3
Summary of Search for Testlet DIF with an Internal Anchor

| Model | -2Loglikelihood | # of Free Parameters | Difference $\chi 2$ | *df* | *P* |
|---|---|---|---|---|---|
| Unconstrained | 8412 | 35 | | | |
| No DIF | 8620 | 18 | 208 | 17 | <.001 |
| Just I DIF | 8511 | 22 | 99 | 13 | <.001 |
| Just II DIF | 8466 | 23 | 54 | 12 | <.001 |
| Just III DIF | 8616 | 21 | 204 | 14 | <.001 |
| Just IV DIF | 8617 | 23 | 205 | 12 | <.001 |
| I & II DIF | 8421 | 27 | 9 | 8 | 0.4 |
| I (C) & II DIF | 8425 | 25 | 13 | 10 | 0.2 |

TABLE 4
Summary of Search for Testlet DIF with an External Anchor

| Model | −2Loglikelihood | # of Free Parameters | Difference $\chi 2$ | *df* | *P* |
|---|---|---|---|---|---|
| **Testlet I** | | | | | |
| No DIF | 1256 | 23 | | | |
| DIF in *c*'s only | 1236 | 25 | 20 | 2 | <.0001 |
| DIF in *c*'s and *a*'s | 1234 | 27 | 22 | 4 | <.0001 |
| **Testlet II** | | | | | |
| No DIF | 1396 | 24 | | | |
| DIF | 1312 | 29 | 84 | 5 | <.0001 |
| **Testlet III** | | | | | |
| No DIF | 1206 | 22 | | | |
| DIF | 1204 | 25 | 2 | 3 | 0.68 |
| **Testlet IV** | | | | | |
| No DIF | 1239 | 24 | | | |
| DIF | 1231 | 29 | 8 | 5 | 0.12 |

## Figure Captions

Figure 1. The values of $a_k$ for Testlet I for males, plotted against score-category.

Figure 2 The values of $c_k$ for Testlet I for males, plotted against score-category.

Figure 3. Summary of the search for testlet DIF using an internal anchor. Each model (described in the text) is plotted at its level on an index of model fit (after Bentler & Bonett, 1980), ranging from 0 for the model with no DIF to 100 for the model with DIF for all testlets.

Figure 4. The values of $a_k$ for Testlet I for males and females compared, plotted against score-category.

Figure 5 The values of $c_k$ for Testlet I for males and females compared, plotted against score-category.

Figure 6. Trace lines for the ten response categories for Testlet I; those for the males are in the upper panel and those for the females are in the lower panel. The modes of the trace lines are in the order of the score-group categories 0–9.

Figure 7. Expected score category on Testlet I plotted against proficiency for males and females.

Figure 8. The difference between males and females in expected score category on Testlet I plotted against proficiency.

Figure 9 The difference between males and females in expected score category on Testlet II plotted against proficiency.

Figure 10. Trace lines and parameters for the six external anchor items.

BEST COPY AVAILABLE

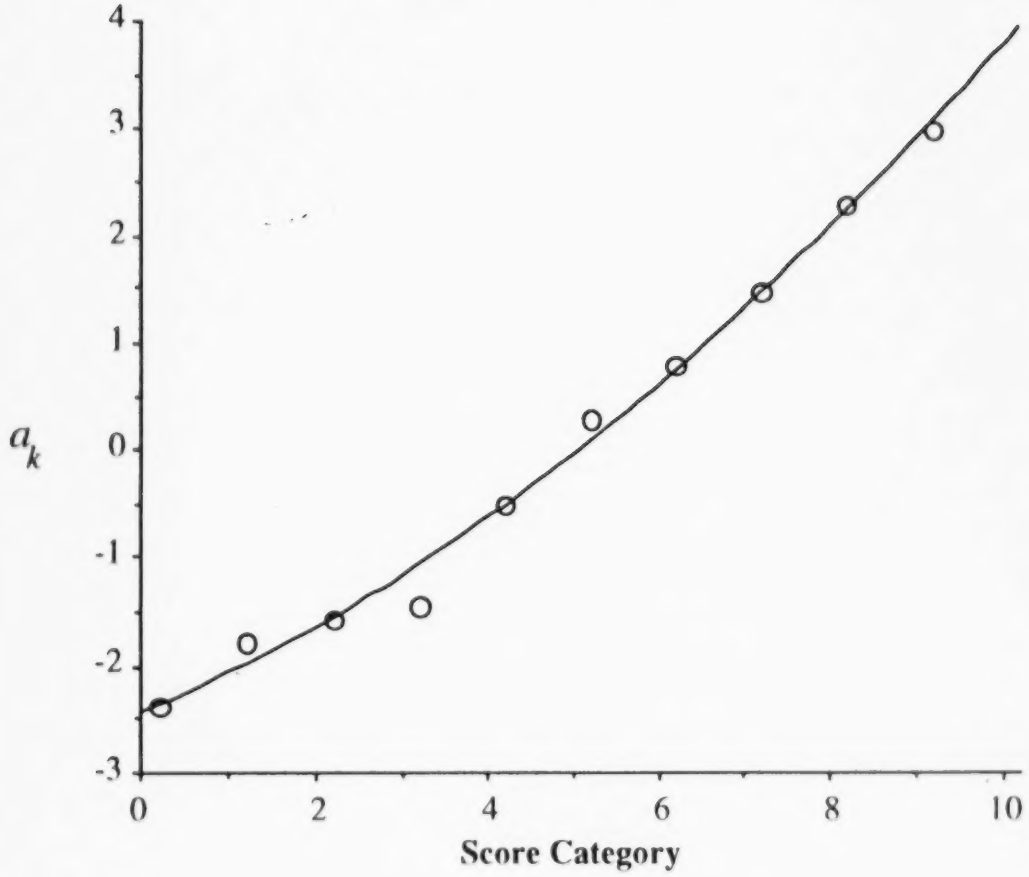


Figure 1

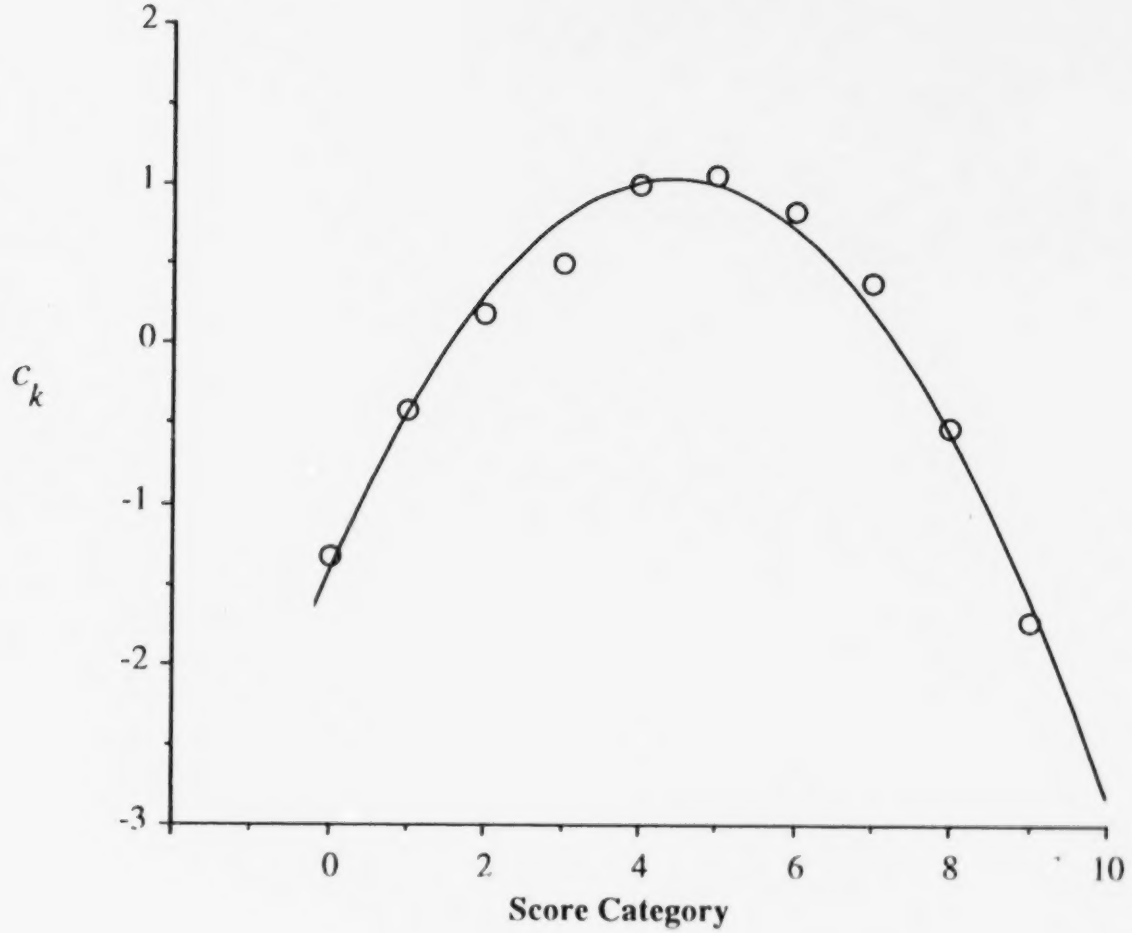


Figure 2

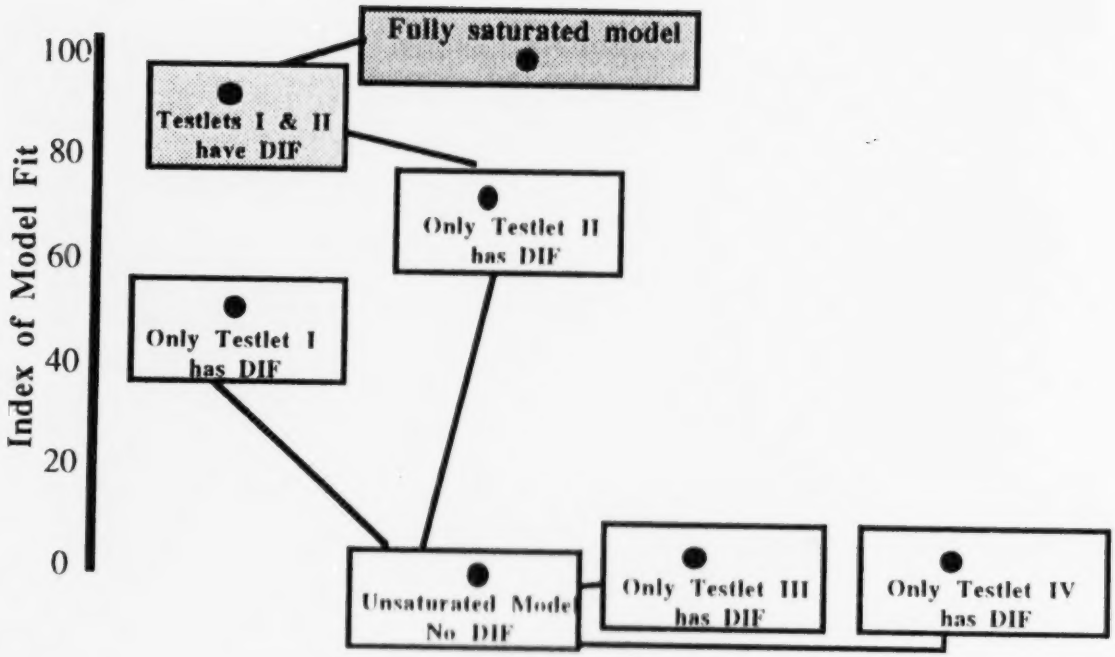
Index of Model Fit
100
80
60
40
20
0
Fully saturated model
Testlets I & II have DIF
Only Testlet II has DIF
Only Testlet I has DIF
Unsaturated Model No DIF
Only Testlet III has DIF
Only Testlet IV has DIF

Figure 3

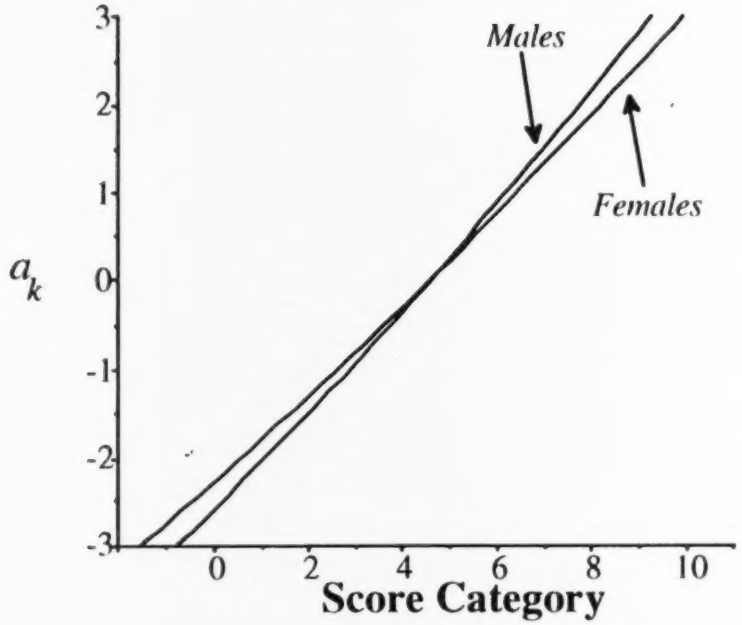


Figure 4

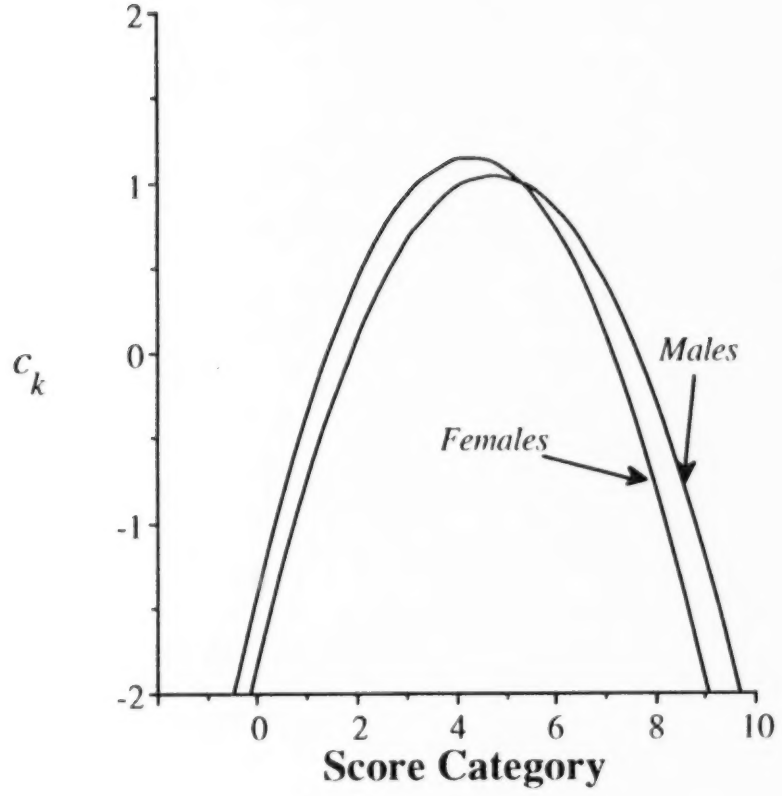


Figure 5

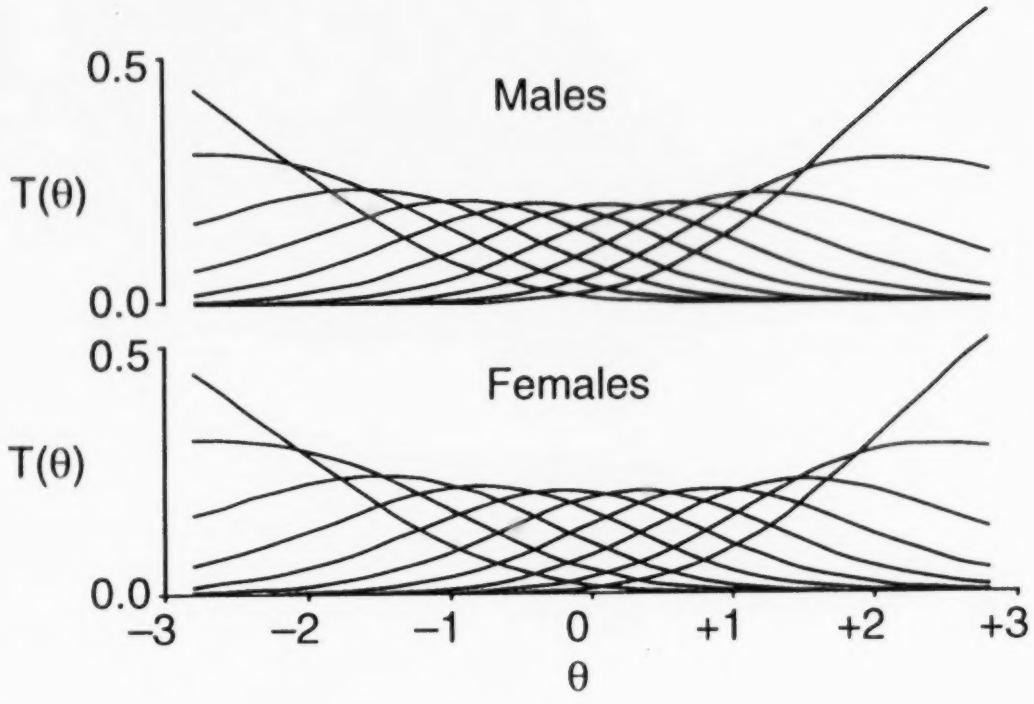


Figure 6

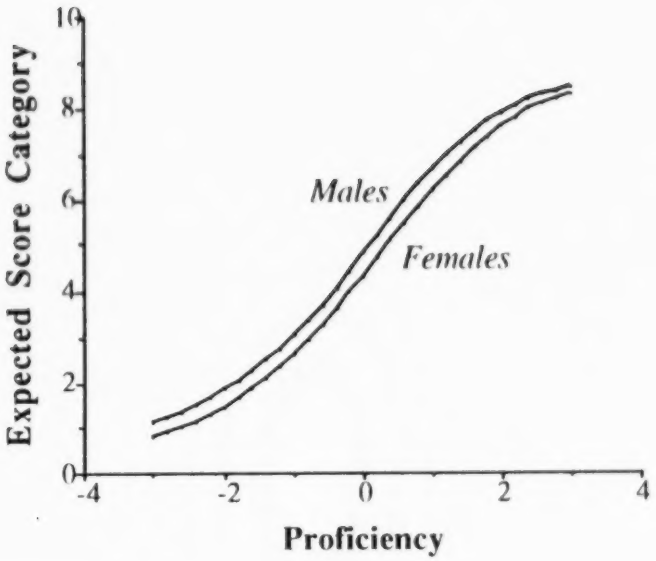
Expected Score Category
10
8
6
4
2
0
-4
-2
0
2
4
Proficiency
Males
Females

Figure 7

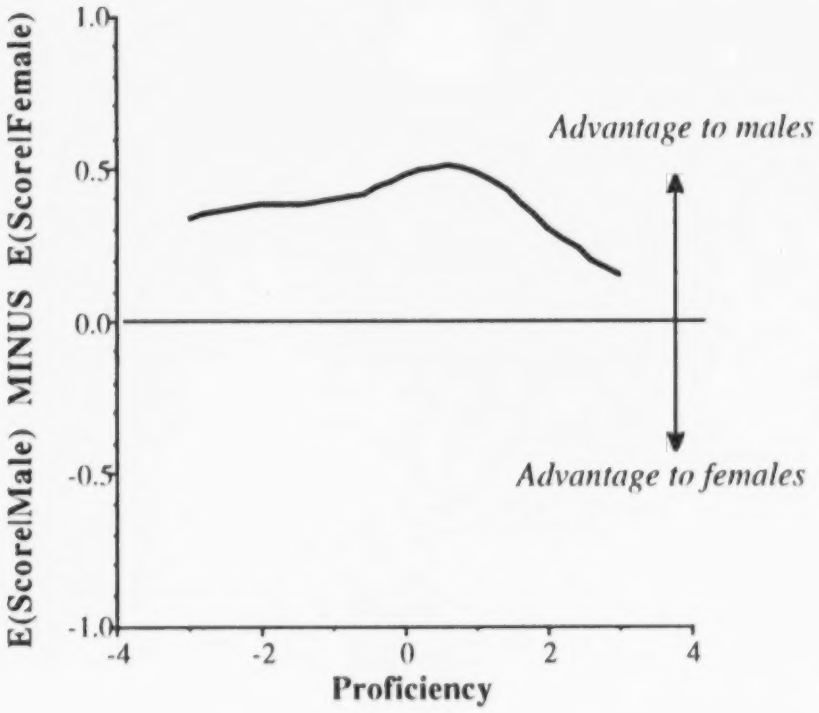
E(Score|Male) MINUS E(Score|Female)
1.0
0.5
0.0
-0.5
-1.0
-4
-2
0
2
4
Proficiency
Advantage to males
Advantage to females

Figure 8

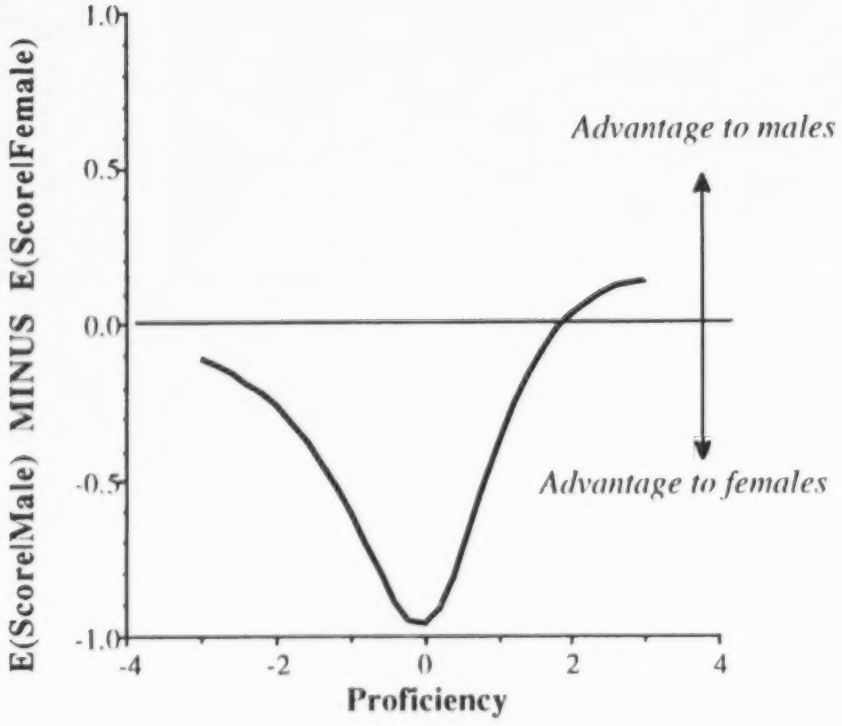
E(Score|Male) MINUS E(Score|Female)
1.0
0.5
0.0
-0.5
-1.0
-4
-2
0
2
4
Proficiency
Advantage to males
Advantage to females

Figure 9

*Parameters of External Anchor Items*

| Item | a | b | c |
|---|---|---|---|
| 1 | 1.4 | 1.7 | 0.1 |
| 2 | 1.4 | 1.3 | 0.1 |
| 3 | 0.8 | 0.6 | 0.2 |
| 4 | 1.0 | -0.5 | 0.1 |
| 5 | 0.8 | -0.4 | 0.2 |
| 6 | 0.7 | -1.4 | 0.2 |

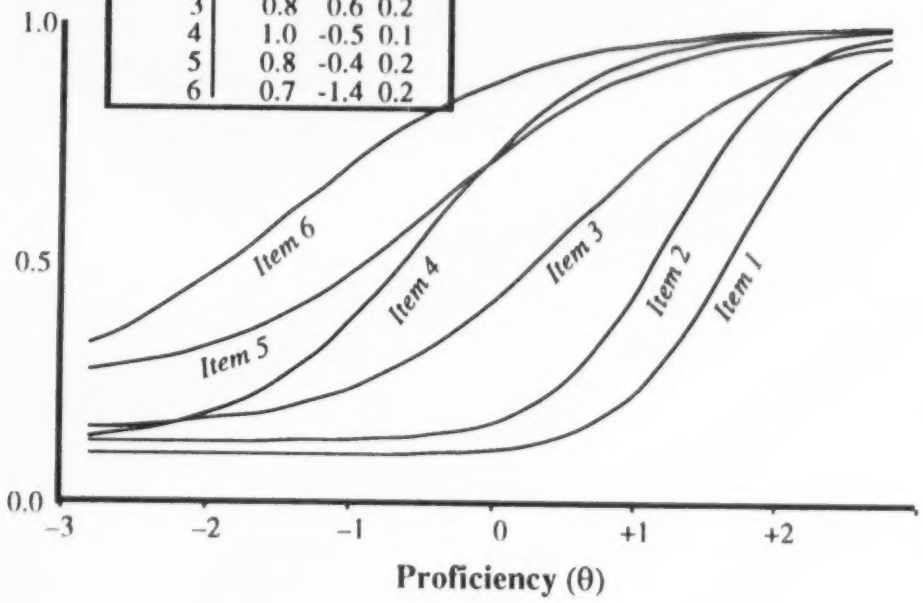




Figure 10